Reviewing the Evidence on Assessing Risk for Child Abuse and Neglect

Della Knoke, MA
Nico Trocmé, PhD

Risk assessment is a central component of crisis intervention in all aspects of practice. Nowhere is this more pronounced however, than in the assessment of families in crisis and children at risk of abuse. Structured risk assessment instruments are promoted to manage increasing demands for child welfare services by providing a mechanism to guide decision making regarding the type and intensity of services required to protect children from subsequent harm. The value of the structured risk assessment instruments is hypothesized to lie in improved consistency and accuracy of workers’ judgments. However, risk assessment models were frequently implemented with little empirical evaluation. Postimplementation studies indicate that many commonly used risk assessment tools fail to attain adequate levels of reliability and validity. A number of challenges to validation have been identified. A more systematic approach to the development and testing of risk assessment instruments is required to support child welfare practice. [Brief Treatment and Crisis Intervention 5:310–327 (2005)]

KEY WORDS: risk assessment, child welfare, review.

Interest in the early identification of children at risk for severe maltreatment emerged in the 1950s. With the advent of radiology, evidence of previous bone fractures in different stages of healing could be detected in children presenting to emergency with trauma. Some of these children died of head injuries after being discharged from hospital in their parents care. Based on this pattern of serious and recurrent injuries, radiologists advocated the removal of these children from their homes to prevent subsequent trauma. For example, one radiologist, John Caffey, noted in 1957 that “correct early diagnosis . . . may be life saving to some of these otherwise helpless youngsters, or it may prevent crippling injuries to others” (cited in Williams & Money, 1980, p. 283). Similar sentiments have been expressed by prominent child welfare advocates. Melton (1991) wrote that respect for the dignity of children implies
a duty “to prevent assaults on that integrity whenever possible” and that “when suffering is foreseeable, its occurrence is preventable, and no action is taken, surely authorities are culpably negligent” (p. 347).

The primary objective of risk assessment processes is to identify, from cases referred to child welfare authorities, the subgroup of children at high risk for future abuse or neglect so that action may be taken to prevent it (Fluke et al., 1995; Pecora, 1991). A second prediction relates to the likely severity of subsequent maltreatment (e.g., Miller, Williams, English, & Olmstead, 1987; Wells, 1985, cited in English & Pecora, 1994). The present paper has two primary purposes. First, conceptual models of structured risk assessment in child welfare practice are examined. Second, validation research is considered to assess the extent to which these instruments are successful in improving consistency and accuracy in assessments of risk for child maltreatment.

The assessment of risk is essential to crisis intervention (Lewis & Roberts, 2001; Roberts, 2000). Child welfare workers traditionally relied on clinical experience in making determinations about which children are likely to experience future maltreatment and the level of intervention required to protect them. The implementation of structured risk assessment in child welfare is fairly recent. In the mid-1980s shifts toward structured risk assessment were documented in the United Kingdom, Australia, the United States, and Canada (Waldfogel, 1998). For example, 26 of the 44 jurisdictions that responded to an American Public Association Survey in 1996 implemented formal/structured risk assessment processes after 1987 (Tatara, 1996).

Structured risk assessment provides a uniform approach and criteria for determining risk. It makes the process used to calculate risk more explicit and assists practitioners in staying focused on the construct being assessed (Lewis & Roberts, 2001). The kind of information collected in risk assessment relates specifically to the likelihood of future harm and is more narrowly focused than comprehensive child and family assessments. These instruments do not necessarily provide sufficient information to address broader aspects of child well-being. They are intended to categorize families on the basis of their likelihood of future maltreatment in order to assist agencies in targeting the most extensive services to children with the highest risk (English & Pecora, 1994; Johnson & L'Esperance, 1984). Structured risk assessment instruments are also designed to guide clinical judgment and facilitate rational approaches to decision making by identifying targets for intervention (English & Pecora, 1994). Thus, these instruments perform circumscribed functions and should be conceptualized as one component of an overall assessment.

Risk assessment is related, but not identical, to safety assessment. Safety assessment focuses on the risk of severe harm in the immediate future so that short-term decisions can be made. For example, emergency removal of a child may be required until the family situation is stabilized and a more complete assessment can be conducted. Safety assessments provide an index of the extent to which an acute family crisis places the child at significant risk for harm. Structured risk assessment tools, on the other hand, are designed to estimate the likelihood that family crisis and maltreatment (the hazard) will reoccur over the longer term. Within a crisis prevention framework, risk assessment facilitates clinical decision making about the interventions required to avert future crises.

Two sets of factors contributed to the widespread implementation of structured risk assessment in child welfare: (a) increasing demand for services, coupled with resource limitations and (b) concerns regarding professional credibility and clinical decision making. Rates of reported child abuse and neglect
have been increasing in Canada and across the western world (English & Pecora, 1994; Trocmé, Fallon, MacLaurin, & Copp, 2002; Waldfogel, 1998; Weedon, Torti, & Zunder, 1988). Broadening definitions of maltreatment may contribute to the increase in referrals. For example, the inclusion of exposure to domestic violence as a reportable form of maltreatment in some jurisdictions has led to significant increases in reports to child protective services (e.g., in Minnesota, as cited in Edelson, 2004). To cope with increased caseloads, increasingly severe cases of abuse and neglect (English & Pecora, 1994; Sedlak & Broadhurst, 1996) and insufficient resources to provide services to all families that may benefit from them (Doueck, English, DePanfilis, & Moote, 1993; Johnson & Clancy, 1988), risk assessment is promoted to manage service demand and to allocate limited resources more effectively (Camasso & Jagannathan, 2000; DePanfilis & Zuravin, 2001; English & Pecora, 1994; Johnson & L’Esperance, 1984).

The second impetus for the adoption of structured risk assessment was concern regarding the validity and reliability of worker judgments about risk. Accuracy in assessing risk is crucial to ensure that appropriate intervention is provided. Structured risk assessment tools provide a systematic way for documenting risk and a framework for making decisions about the type and intensity of services required. Clinical experience and “practice wisdom” have traditionally informed workers’ service decisions. Good clinical skills are an integral part of the assessment process. However, their sufficiency in making determinations about the likelihood of future severe maltreatment has been questioned (Jagannathan & Camasso, 1996; Rossi, Schuerman, & Budde, 1996; Weedon et al., 1988). For example, early research by Margaret Blenkner (1954, as cited in Baird, Wagner, Healy, & Johnson, 1999) found poor prediction of case outcomes among expert social workers. In addition, child welfare experts and workers from four states in the United States showed significant variability in their judgments about whether each of 70 case vignettes should be opened for in-home services and whether family preservation services or out-of-home care was recommended (Rossi et al., 1996). Differences in clinical skills, practice experiences, educational background and theoretical perspectives influence the kinds of information workers consider relevant (Baird et al., 1999) and the relative weight assigned to case information (Blenkner, 1954, as cited in Baird et al., 1999). The development of practice wisdom takes time and, as new workers replace more experienced ones, differences in the level of clinical expertise may influence service decisions (Cash, 2001). Concerns about the accuracy of clinical judgment were reinforced by highly publicized deaths of children known to child welfare services (Fuller, Wells, & Cotton, 2001; Weedon et al., 1988). Estimates suggest that approximately half of all child abuse fatalities involve children who had been previously reported to child protective services (National Committee to Prevent Child Abuse, 1996, as cited in Besharov, Robinson-Lowry, Pelton, & Weber, 1998). The value of more standardized approaches to risk assessment has long been recognized in other fields. For example, research in criminology consistently reveals that clinical judgment is less accurate in predicting future behavior than judgments based on structured risk assessment instruments (e.g. Dawes, Faust, & Meehl, 1989; Meehl, 1954).

**Types of Risk Assessment Instruments**

In general, risk assessment models are divided into two primary types: actuarial based and consensus based. Actuarial-based instruments incorporate client characteristics shown to be statistically predictive of future maltreatment
(Rycus & Hughes, 2003). The relative weight that each risk factor contributes to the overall risk rating is determined by a formula designed to maximize predictive accuracy. Assessments are scored in a mechanical manner to ensure consistency across cases and settings. The use of actuarial tools and actuarial approaches to risk assessment must be differentiated. The classification of risk assessment approaches as actuarial requires both the use of a tool that is developed to maximize predictive accuracy based on statistical relationships and the use of risk classifications derived solely on the basis of these instruments. That is, adherence to an actuarial approach requires that the risk classifications generated and the weight assigned to individual items are not modified by clinical judgment (i.e., by consideration of case characteristics other than those specified by the risk assessment instrument).

Consensus-based instruments, on the other hand, use expert clinical judgment to determine which client characteristics should be assessed. In the selection of variables, expert judgment is informed by both clinical experience and knowledge of the research literature. In contrast to actuarial models’ formulaic approach to assigning an overall risk level, consensus-based approaches guide decision making (Marshall & English, 2000). Workers consider risk factors and arrive at a professional judgment about the overall level of risk. Many of the consensus-based instruments may also be referred to as “blended” instruments. Blended instruments are those developed through nonempirical processes but later subjected to empirical validation (National Resource Centre on Child Abuse and Neglect, 1984, as cited in Cash, 2001).

What Are the Risks in Child Maltreatment?

Child maltreatment includes a heterogeneous group of acts (of commission and omission) that are generally categorized into four primary forms: physical abuse, neglect, sexual abuse, and emotional abuse. The Canadian Incidence Study of Reported Child Abuse and Neglect (CIS-98) (Trocme´ et al., 2001), the first nationally representative study of its kind in Canada, found that neglect and physical abuse were the most commonly investigated primary forms of child maltreatment (40% and 31%, respectively). Sexual abuse and emotional abuse accounted for 10% and 19% of all cases, respectively. The CIS-98 indicated that most cases investigated for child maltreatment involve no harm or minimal harm (i.e., some harm but no treatment required). Across all categories of maltreatment, 13% of children sustained some physical harm. Three percent had physical injuries that required medical treatment. Other studies of child welfare samples in the United States (e.g., Zuravin, Orme, & Hegar, 1995) and the United Kingdom (e.g., Creighton, 1985) have found that severe physical harm occurs in a minority of maltreated children. Findings from national community surveys of family violence in the United States also indicate that although the use of physical force against children is prevalent, acts that have a high probability of resulting in severe physical harm (i.e., beatings, threatening with a weapon) occur at much lower frequencies than milder forms of abuse (Gelles & Straus, 1987, 1988).

The CIS-98 also suggests that the likelihood of harm and severe harm varies across forms of maltreatment (e.g., Trocme´, MacMillan, Fallon, & DeMarco, 2003). Physical abuse was more likely than the other primary forms of maltreatment to result in some form of physical injury (28%) and severe physical harm (4%). In addition, within the broad categories of maltreatment, the likelihood of harm may vary considerably. For example, among children who had substantiated physical abuse, almost three quarters investigated for shaken baby syndrome sustained physical harm,
with approximately half resulting in severe harm. In contrast, a large proportion of cases involving inappropriate punishment and “other” physical abuse resulted in no harm (63% and 46%, respectively). Three percent of children subjected to inappropriate punishment and 10% of other physical abuse resulted in physical harm requiring medical attention. Thus, in general, a minority of the children who come to the attention of child welfare services may be at risk for maltreatment that results in severe harm. However, some forms of maltreatment and abuse contexts may place children at greater risk than others.

**Criticism/Concerns About Risk Assessment in Child Welfare**

Since the implementation of structured risk assessment in child welfare, several concerns and cautions have been expressed about the scant theoretical and empirical support for these instruments (Cicchinelli, 1991; English & Pecora, 1994; McDonald & Marks, 1991; Pecora, 1991; Wald & Woolverton, 1990). The most fundamental concerns relate to the psychometric properties of commonly used instruments. Most risk assessment models were developed and implemented with little or no research to establish validity or reliability (English, Aubin, Fine, & Pecora, 1993; McDonald & Marks, 1991; Pecora, 1991; Wald & Woolverton, 1990) and with little, if any, empirical testing (McDonald & Marks, 1991; Rycus & Hughes, 2003). Several threats to reliability and validity have been identified. First, the factors incorporated in risk assessment instruments were frequently selected on the basis of research studies differentiating maltreating from nonmaltreating families (Wald & Woolverton, 1990). However, many of these factors are present in a large proportion of families who come to the attention of child welfare services. The fundamental question in child welfare contexts is among families who have or who are likely to have maltreated their children, what factors influence the likelihood that maltreatment will continue? Second, instruments are often modified over time or when adopted for use in other jurisdictions without consideration for the effect of these modifications on the psychometric properties of instruments (McDonald & Marks, 1991; Rycus & Hughes, 2003). Third, they are used for a variety of purposes for which they are not intended and for which there is no empirical support (e.g., whether it is safe to return a child from foster care to the family home) (Wald & Woolverton, 1990). Finally, workers were often provided with little or no training to ensure that instruments are being implemented as intended (e.g., Cicchinelli & Keller, 1990; Doueck et al., 1993; Sullivan, 1997). Given these concerns, several factors may compromise the extent to which commonly used risk assessment instruments improve the consistency and accuracy of workers’ judgments of risk. After implementation, there have been numerous research efforts to validate commonly used instruments. These studies are examined below.

**Validation Research**

**Reliability Studies**

Interrater reliability is the form of reliability assessed most frequently. These studies examine the extent to which independent raters using a specific risk assessment instrument agree on the risk level assigned to cases. There are two predominant approaches. First, case information is compiled, summarized, and presented in the form of case vignettes. Raters trained in the use of the risk assessment instruments independently apply the risk assessment criteria and assign a risk rating to each
case. The second method involves the same independent assessment, but case files, rather than vignettes, are used to establish the level of risk. Several early studies were criticized for including too few cases and not including a full range of risk levels (e.g., Nasuti & Pecora, 1993) and for the use of correlations as the index of interrater agreement, without adjusting for agreement that may occur by chance alone (e.g., Allen, 1988; Doueck et al., 1993).

Some positive results have been reported in studies that employ more rigorous methodologies. Camasso and Jagannathan’s examination of the New Jersey Risk Assessment Matrix consensus-based models revealed interrater reliability coefficients in the .85–.90 range (as cited in Jagannathan & Camasso, 1996). Wood (1997) assessed the interrater reliability of an actuarial instrument for 63 randomly selected case files. A median kappa value of 0.66 was attained. Other studies suggest that levels of reliability vary considerably across instruments. For example, Baird et al. (1999) compared the reliability of two consensus-based instruments, the Washington Risk Assessment Matrix (WRM) and the California Family Assessment Factor Analysis (CFAFA), to one actuarial system, the Michigan Structured Decision Making System’s Family Risk Assessment of Abuse and Neglect (FRAAN). Case readers assigned risk ratings to 80 cases, selected from four sites. Superior levels of agreement were attained with the actuarial instrument (FRAAN), both in terms of proportion of cases with 100% agreement and in median kappa values. For example, median kappa values were .245 and .211 for the WRM and CFAFA, respectively, as compared to .635 for FRAAN. To the extent that the low levels of reliability documented in these two consensus-based systems represent the performance of consensus-based approaches in general, significant concerns are raised regarding these commonly used instruments. Also, although the reliability of the FRAAN model was significantly better than the two consensus-based models, the level of reliability attained is lower than desired.

One of the factors that may influence overall reliability is the variability in reliability across risk items. Items that are more concrete tend to be rated more consistently than items that require subjective evaluation (Baird et al., 1999; McDonald & Marks, 1991; Wood, 1997). For example, high levels of interrater reliabilities were attained for items such as child age and number of prior openings. However, lower reliabilities were evident for more subjective items such as whether excessive or inappropriate discipline was used or assessments of whether a caregiver’s emotional instability compromised his/her functioning as a caregiver (Wood, 1997). In addition, Sullivan (1997) found that risk elements that required more extensive clinical observation (i.e., assessment of parent–child interactions, child’s mental health, and development) were rated less consistently across workers than other risk elements. Because risk assessment instruments vary considerably in the number and type of items included (Rycus & Hughes, 2003), the extent to which poor reliability on these items influences overall reliability ratings is expected to vary by instrument.

**Implementation Issues.** Gambrill and Shlonsky (2001) noted that “unnecessary risk may result from the use of invalid instruments as well as the misuse of a valid risk assessment measure” (p. 84). This quote captures the importance of assessing how risk assessment instruments are implemented in practice. To ensure that instruments are being implemented as intended, adequate training and quality assurance programs are required (Barber & Delfabbro, 2003; English & Aubin, 1989; English, Aubin, & Austin, 1992). Numerous researchers have documented the difficulty in implementing risk
assessment models (Cicchinelli & Keller, 1990; DePanfilis, 1996; Doueck et al., 1993; Fluke et al., 1994; Hornby, 1989; Kern, Baumann, & Sheets, 1994; Pecora, 1991). For example, risk assessments may be used to verify or support decisions that workers have already made (Cicchinelli & Keller, 1990; English & Pecora, 1994; Fluke, 1992, as cited in English & Pecora, 1994). In addition, case file audits in three pilot regions in Australia found high levels of noncompliance and data fabrication after implementation of a structured risk, harm, and safety protocol (Barber & Delfabbro, 2003). These practices and the insufficient training provided to workers may result in levels of reliability that are lower in practice settings than in the controlled study settings described above. Poor implementation and resultant low levels of reliability in practice may also lead to erroneous conclusions regarding the relationship between risk ratings and outcomes, thus compromising efforts to validate risk assessment instruments.

Validity Studies

Empirical studies assess two primary forms of validity. Criterion validity has been more extensively examined. A smaller number of studies have examined the construct validity of risk instruments and individual risk items.

Predictive Validity. Efforts to establish criterion (predictive) validity are complicated by a lack of clearly defined goals and objectives (Baumann et al., 1997) and poor implementation of risk assessment instruments in practice (Doueck et al., 1993). The primary intent of risk assessment is the prediction of subsequent maltreatment. Studies typically measure subsequent maltreatment by a new referral to child protective services or by a new substantiated report of maltreatment during a specified period of time following completion of the risk assessment. There is some controversy about which represents the most appropriate criterion. Some have argued that substantiated cases may meet higher evidentiary standards. However, others caution that the decision to substantiate a case is an administrative decision that may not be an accurate index of whether or not maltreatment actually occurred (e.g., Drake, 1996; English, Marshall, Brummel, & Orme, 1999). Although the intent of the risk assessment processes is to identify children at risk for harm, the criteria used in the research only provide an index of whether future maltreatment is alleged or substantiated (i.e., not whether subsequent maltreatment or harm actually occurs).

Risk assessment instruments have traditionally been implemented without adequate evidence to indicate that they accurately identify the children most likely to experience subsequent maltreatment. Little information has been published on the predictive validity of risk assessment tools used in Canada. However, several studies have assessed the validity of instruments that have been implemented in a number of jurisdictions in the United States. Two indices of predictive accuracy are particularly relevant. Although the overall level of prediction accuracy provides a global index of how well the model predicts outcomes, the rates of false positives and false negatives provide important information about where the errors in prediction occur. Value judgments are required to determine the level and type of error that is “reasonable.” Prediction error that is related to the tendency of instruments to miss high-risk cases (false negatives) may result in the inadequate provision of services and tragic consequences for children. When instruments tend to erroneously classify cases as high risk (false positives), intervention may be unnecessarily intrusive and inappropriate to meet child and family needs.
The rate of false positives has been estimated to range from a low of 15% to a high of 83% (Pecora, 1991). In general, however, studies report that between 14% and 29% of children are erroneously identified as likely to experience future maltreatment (Baird, 1988; English, Marshall, Brummel, & Coghlan, 1998; Johnson & L’Esperance, 1984; Marks & McDonald, 1989; Weeden et al., 1988). The rate of false negatives also varies considerably across instruments. In a review of risk assessment research, Lyons, Doueck, and Wodarski (1996) found that between 14% (Washington) and 66% (Vermont) of cases re-referred were not previously identified as cases in which subsequent maltreatment was considered likely. Other evaluations of the WRM indicated false-negative rates of 14–22% (English et al., 1998). Johnson and L’Esperance found that between 25% and 30% of children were erroneously classified as not likely to experience future abuse. Studies of other jurisdictions also indicate false-negative rates of 22–45% (Baird, 1988; Baird, Wagner, & Neuenfeldt, 1993; Lyons et al., 1996; Marks & McDonald, 1989). Although the high rate of false negatives appears to be atypically high in the Vermont study, the findings indicate that, depending upon the instrument used, between one in six and two in three children likely to experience future harm may be missed using current risk assessment instruments.

Overall, the predictive performance of risk assessment models has been disappointing (Baird & Wagner, 2000; Camasso & Jagannathan, 2000; Rittner, 2002). In general, studies find that less than a third of the variance in maltreatment recurrence is explained by the factors included in risk assessment instruments (Baird & Wagner, 2000; Camasso & Jagannathan, 1995; Fuller et al., 2001; Rittner, 2002). Difficulty in explaining and predicting future behavior is not specific to child maltreatment. It is not uncommon to find similarly low levels of explained variance in predictive models developed to explain phenomena such as criminal recidivism and cancer recurrence (Baird & Wagner, 2000). Like child maltreatment, these phenomena are multidetermined. In child welfare, initial assessments provide a snapshot of child, caregiver, and family characteristics. Some of the factors associated with future maltreatment may be distal and relatively invariant (e.g., parent history of abuse as a child). Others are dynamic and may change over time (i.e., single-parent status, number of children in the family, living arrangements), influencing the likelihood of maltreatment recurring.

Most instruments examine the risk for any subsequent maltreatment. However, some studies suggest that risk assessment may be more valid for some forms of maltreatment than others (e.g., English et al., 1998; Wood, 1997) and that the nature of the prediction errors may depend on maltreatment type (Johnson & Clancy, 1989). For example, using a risk matrix, Johnson and Clancy (1989) found twice the rate of false positives among cases of physical abuse (33%) than sexual abuse (15%). In addition, the rate of false negatives was 24% for physical abuse and 41% for sexual abuse. This finding is consistent with studies that suggest that different forms of maltreatment may be etiologically distinct (e.g., Hamilton & Browne, 1999; Swanston et al., 2002; Wood, 1997). It also suggests that validation studies may be compromised by variability in the predictive accuracy of instruments across forms of maltreatment.

Baird and Wagner (2000) argued that the viability of precise prediction is particularly difficult for phenomena with low base rates. The extent to which maltreatment recurrence represents a low base rate event is debatable. Rates of maltreatment recurrence are estimated to vary from 18% (Doueck et al., 1993) to 60% (Wolock, Sherman, Feldman, & Metzger, 2001), depending on the operational definition of recurrence adopted (report vs. substantiated report) and the time frame over which cases are
monitored (months vs. several years). However, if the criterion of interest is the severe harm, the low base rate argument applies. For example, as discussed earlier, the CIS-98 indicated that physical and emotional harm requiring treatment was evident in 3% and 15% of children investigated for maltreatment (Trocme et al., 2001). Baird and Wagner propose shifting the focus from predicting who will “fail” and who will not “fail” to assigning cases to risk categories based on “observed rates of behavior.” That is, individuals in higher risk groups are conceptualized as requiring more services “because cases in this designation tend to fail at higher rates than cases in other classifications” (p. 852). Classification in a high-risk category is not a prediction that subsequent maltreatment will occur. It indicates that given a certain cluster of risk factors, the likelihood of subsequent maltreatment is higher in this than in other risk categories.

Several studies indicate that risk designations are associated with different rates of subsequent maltreatment. An evaluation of the WRM found that cases rated “low risk” had lower rates of re-referral than cases designated “moderate to high risk” (18% vs. 26%) (English et al., 1993). In an evaluation of the Alaska Risk Assessment model, 83% of very high risk cases of abuse (physical, emotional, and sexual) not removed from the home were subsequently abused, as compared to only 3.3% cases classified as very low risk (Baird, 1988). The same study found that 70% of high-risk neglect cases (physical, medical and emotional) not removed from their homes had subsequent neglect reported, as compared to 7.5% of neglect cases classified as low risk. Wood (1997) found that 52% and 34% of abuse cases classified as high to very high risk had new allegations or new substantiations, respectively, during a 2-year follow-up. In contrast, 12% and 5% of the low-risk cases had new allegations or new substantiations. The rate of new allegations or substantiated reports in cases of neglect was 45% and 19% over the same period of time, as compared to 4% and 1% of the low-risk cases. Finally, using the Vermont Family Risk Assessment Matrix, 61% of the families rated high risk had a subsequent founded report, as compared to 36% and 24% of the moderate- and low-risk groups, respectively (Weedon et al., 1988). In general, these studies suggest that the classification of cases by risk level assists in the identification of subgroups with differential probability of subsequent maltreatment. However, the accuracy of risk assessment instruments in differentiating groups at lower as compared to higher risk for future maltreatment varies across instruments. For example, 5% of the low-risk cases in Wood’s study as compared to 24% of the low-risk cases in the Weedon et al. study had subsequent founded allegations.

Baird and Wagner (2000) suggested that there may be significant differences in the predictive validity of consensus- and actuarial-based instruments. Their study examined two consensus-based instruments (WRM and the CFAFA) and one actuarial model used in Michigan (FRAAN). The three instruments were compared on their ability to predict new investigations and new substantiations 18 months following the risk assessment. A total of 1,400 cases from four states in the United States were classified as low, medium, or high risk based on case reader assessments. The distributions of these variables across the three risk categories for each of the three models were analyzed using proportional difference testing. The actuarial model produced substantially better risk classifications than either of the consensus-based approaches. Using the actuarial approach, the proportion of cases with new investigations and new substantiations increased as the risk level increased. In addition, this model accurately classified as high risk the largest proportion of cases with new investigations or substantiations.
In contrast, cases classified as moderate and high risk using consensus-based models did not differ in on either index of subsequent maltreatment. Baird and Wagner also computed a Dispersion Index for Risk (Silver & Banks, 1998, as cited in Baird & Wagner, 2000) to assess the extent to which rates of reabuse for each level of risk differed from the rate for the sample as a whole. Again, the actuarial approach was superior to consensus-based approaches in predicting the incidence of subsequent investigations and maltreatment substantiations.

**Construct Validity.** Two aspects of construct validity are examined in the literature. Several studies examine the association between risk level and service response. A few studies examine the relationship between individual risk items and validated measures of the same constructs.

One objective of structured risk assessment models is to create a systematic and structured approach to service allocation. In theory, the highest risk cases should receive more services or more intense services. Risk level should, therefore, be associated with differences in services. The research findings are mixed. Marks and McDonald (1989) found a high correlation between risk level and service intensity. Other studies found that high ratings on particular risk items are associated with distinctive service patterns (Fluke, 1991; Jagannathan & Camasso, 1996; Norlin, 1994). These findings suggest that risk assessment information is being used to guide decisions about appropriate interventions. That is, services are being allocated to ameliorate the specific conditions that increase risk. However, other studies suggest that there is little relationship between risk level and the type or intensity of services (Fluke et al., 1994; Johnson & L’Esperance, 1984). For example, using a multiwave panel design Camasso and Jagannathan (2000) examined the relationship between risk level (WRM) and service intensity for 395 cases. Each case had three or four assessments over a 1-year period. The risk assessment level was found to be unrelated to the service intensity, as measured by the number of service contacts in the time between risk assessments.

Placement in foster care represents one of the most intensive service responses. Using bivariate analyses, higher risk on the WRM was associated with greater likelihood of subsequent placement (English et al., 1993). Baird and Wagner (2000) also found that new placement rates were higher among the high-risk than among the moderate and low-risk groups for two consensus-based models (WRM and CFAFA) and one actuarial model (FRAAN). The overall risk score from the Ontario Risk Assessment Tool was a significant predictor of placement status and increased accuracy for the classification of children placed in out-of-home care, compared to the proportion of correct classifications based on chance (Leschied, Chiodo, Whitehead, Hurlkey, & Marshall, 2003). In addition, a subset of risk factors that included family ability to cope with stress, availability of social supports, caregiver motivation, and caregiver cooperation with investigation differentiated children placed from the children who were not placed.

Few studies have examined the extent to which individual risk assessment items measure what they purport to measure. Two studies suggest poor correlations between risk assessment items and validated measures of the same constructs. The study by Kolko (1998) of the Pittsburgh Service Delivery model found no consistent association between risk assessment measures performed on 90 families and clinical measures of child, parent, and family functioning. English and Graham (2000) examined the correlation between workers ratings of risk on nine risk factors of the WRM and independent measures of the same risk constructs. Significant
correlations were evident for four of the five caregiver risk factors including indices of physical, mental, and emotional impairment. However, low correlations were found for child risk factors associated with developmental or behavioral concerns and socioeconomic factors including stress and social support. Thus, though risk assessment items may have face validity, it is unclear what many of these items are measuring. The poor construct validity of individual items may be related to poor reliability or to the inclusion of an insufficient number of items to adequately assess constructs such as child behavior concerns. If individual risk items are being used to define areas for interventions (e.g., substance abuse issue), it is crucial that workers are able to ascertain whether or not a particular concern is present with accuracy (Table 1).

### Discussion

It would be remiss to discuss risk assessment instruments without acknowledging the concerns raised by many child welfare advocates. Structured risk assessment models have been criticized for their focus on child physical safety and risk reduction to prevent future maltreatment, without adequate attention to child emotional and cognitive well-being (Anglin, 2002; Gambrill & Shlonsky, 2001; Trocme & Chamberland, 2003). Critics argue that insufficient attention is given to the needs of the majority of maltreated children who are not likely to be physically endangered but who are, nonetheless, at risk for a variety of long-term social, emotional, and behavioral problems. Preventing future maltreatment is given primacy in decision making and the allocation of services, with insufficient services allocated to the treatment of children and youth for whom maltreatment that has already occurred (Anglin, 2002). Service allocation on this basis fails to address the needs of a significant proportion of children referred to child protection services. This approach is reinforced by research that limits definitions of the hazards to the occurrence of future maltreatment. It does not predict which children are at greatest risk for adverse emotional and developmental consequences. Thus, the focus on the level of risk for subsequent maltreatment fails to capture the different kinds of risks maltreated children face.

It is important to note that the intended purpose of current instruments is the identification of children at risk for subsequent maltreatment.

### TABLE 1. Challenges to the Validation of Structured Risk Assessment Tools

<table>
<thead>
<tr>
<th>Reliability</th>
<th>Insufficient evidence to indicate adequate interrater reliability in studies using expert raters. Poor implementation fidelity may compromise the reliability of instruments in practice.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>Poor predictive accuracy, with little variance in maltreatment recurrence accounted for by risk assessment items. Some evidence that criterion validity is better if classification accuracy is used as the criterion. However, there is significant variability across instruments. Mixed evidence for the construct validity of risk assessment instruments with regard to the relationship between risk level and service patterns or intensity. Poor construct validity of individual risk items. Elements requiring subjective evaluation or more extensive clinical observation are identified as particularly problematic.</td>
</tr>
<tr>
<td>Approaches to validation</td>
<td>The definition of any future maltreatment as the hazard to predict identifies the average effect of these factors across all forms of maltreatment and levels of severity. Subgroup variability may compromise validation studies.</td>
</tr>
</tbody>
</table>
This does not preclude the development or use of instruments to predict other outcomes of interest. In addition, the above objections are not inherent properties of the instruments themselves. They relate more specifically to changes in the orientation and organization of child welfare practice that occurred as a consequence of adopting the “risk perspective” (Anglin, 2002, p. 243). The primary issue examined in the current paper is how well do risk assessment instruments in child welfare meet their objective—to accurately identify the subset of children at risk for future harm.

In their seminal paper, Wald and Woolverton (1990) identified three main characteristics of good risk assessment instruments in child welfare. First, valid instruments should be able to differentiate cases in which subsequent maltreatment is likely from those in which it is not. Although this criterion calls for predictive accuracy, it may also be sufficient that instruments categorize, with the highest degree of accuracy possible, cases that differ in their relative likelihood of subsequent maltreatment so that services may be allocated to prevent it. Second, the factors used to assess risk must be measured as accurately as possible. That is, individual risk items must have construct validity and must be measured reliably. Finally, instruments should be able to determine the risk of subsequent maltreatment in the context of specific interventions. Available research suggests that currently used risk assessment tools do not satisfy these criteria.

The interrater reliability of risk assessment instruments was moderate to good in some studies. Other studies suggest significant variability in the way case information is rated across workers. For example, although the reliability of the FRAAN, an actuarial tool, was superior to its two consensus-based comparisons, all instruments performed below levels required to establish a desirable level of consistency across workers. These estimates of reliability are derived from studies in more controlled settings than are found in practice and are likely to be optimistically biased. Raters are highly trained and are given limited and uniform information. In practice, numerous contacts with families and collateral contacts provide multiple opportunities for observation and multiple sources of information. Given the implementation challenges documented, estimates of reliability attained in practice settings are likely to be lower. Poor agreement regarding whether a case is high risk is likely to lead to disparities in the nature and intensity of services implemented.

In general, studies suggest that the predictive accuracy of risk assessment instruments is limited. In particular, the false-negative rates raise questions about overreliance on these instruments. Baird and Wagner (2000) suggested that, given the nature of the phenomenon, an adequate level of precision may be difficult to attain. Several studies suggest that risk assessment instruments assist in categorizing cases in a way that captures differential probabilities of subsequent maltreatment. However, there is significant variability across instruments, and error rates are too high to provide sufficient confidence that children at high risk will not be missed. Some errors may, as Baird and Wagner suggest, be related to the difficulties inherent in predicting complex and multidetermined behaviors. However, it is important to ascertain how much of the error in classification is attributed to dynamic and multifaceted nature of child maltreatment and how much is attributed to measurement error resulting from instruments that are poorly conceptualized, poorly constructed, and poorly implemented.

Evaluations of construct validity were mixed. Risk assessment instruments are implemented with the objective of improving the effective allocation of services to cases with the highest risk for subsequent maltreatment. Risk level is
expected to guide decision making about the services required. Some studies revealed associations between risk level and service intensity. For example, higher levels of placement were consistently found among high-risk cases. Other studies indicate that high levels of risk on individual items or particular clusters of items are associated with particular patterns and intensity of intervention. Some studies found little concordance. Thus, the extent to which the level of risk guides decisions regarding the nature or intensity of service is unclear. The construct validity of individual risk items was found to be poor in two studies. Given that individual items (e.g. whether the caregiver has a drug or alcohol problem) may guide the selection of interventions, effort must be taken to improve the accuracy with which risk constructs are measured.

Levels of reliability and validity vary considerably across risk assessment instruments. This is not surprising given vast differences in the number and type of risk criteria included (Cicchinelli & Keller, 1990; McDonald & Marks, 1991) and in approaches to aggregate risk information (Cicchinelli & Keller, 1990). Across jurisdictions, the number of items included in risk assessment instruments varies from 6 to over 40 (Rycus & Hughes, 2003), with approximately 40% of criteria unique to a single model (Lyons et al., 1996). Thus, there is clearly no consensus about what elements are required to accurately assess risk. Some of the evidence suggests that actuarial models exceed consensus-based models on indices of reliability and validity. A more systematic comparison of these approaches is warranted to examine the extent to which superior psychometric properties of the FRAAN represent characteristics of actuarial approaches in general.

In general, risk assessments specify the conditions under which maltreatment is likely to reoccur. The influence of a variety of child, caregiver and family characteristics are considered in this process. However, with the exception of the study of Camasso and Jagannathan (2000), other studies do not incorporate the potential influence of child welfare intervention on the likelihood of future maltreatment. This omission confounds interpretations of the relationship between risk factors and maltreatment recurrences and may influence the findings of predictive validity studies (Doueck, Bronson, & Levine, 1992; Wald & Woolverton, 1990). For example, Camasso and Jagannathan’s findings suggest that services reduce the incidence of subsequent reports of abuse and neglect. Thus, treatment effects represent significant intervening variables that should be examined as potential determinants of maltreatment recurrence.

There are several additional limitations to current validation studies. First, studies that include a range of maltreatment forms such as neglect, and physical, sexual, and emotional abuse tend to collapse maltreatment forms, assuming a generic set of risk factors. However, some research suggests that different etiological factors underlie different forms (Hamilton & Browne, 1999; Swanston et al., 2002; Wood, 1997). Baird (1988) found that some risk factors were related to abuse but not to neglect. Thus, the use of a generic set of risk factors may neglect predictors that are important for some forms of maltreatment. In addition, variability in predictive accuracy across forms of maltreatment may compromise efforts to validate risk assessment instruments. Second, studies utilize new reports or new substantiations as the criterion to establish predictive validity. There is no empirical evidence to indicate that risk assessment tools are effective in identifying the children who are at risk for subsequent harm. The answer to the question “at risk for what?” has significant implications for the type of service response warranted. Having knowledge that a child is likely to experience maltreatment in the future undoubtedly has an influence on
whether services are provided. However, this knowledge provides little information regarding the likely severity of that maltreatment or about the nature or intensity of services required to prevent it. Third, studies have been criticized for paying insufficient attention to the influence of strengths that may modify the influence of risk factors (Wald & Woolverton, 1990).

A commitment to evaluating risk assessment tools is evident in the research literature. Unfortunately, these evaluations are being conducted often years after formal implementation of risk assessment models. One and one half decades after the paper of Wald and Woolverton (1990), concerns about the reliability and validity of commonly used instruments persist (Camasso & Jagannathan, 2000; Rycus & Hughes, 2003). Although structured risk assessment has been shown in other fields to hold promise (Dawes et al., 1989; Grove & Meehl, 1996), a more extensive and systematic approach to the development and testing of child maltreatment risk assessment tools is needed to support child welfare practice.

Implications for Practice

At this stage in their development, caution is warranted in using risk assessment tools in child welfare practice. These tools can help to structure assessments and to serve as a checklist of factors to be considered. However, the research suggests that reliance on many commonly used instruments as the only basis for determining level of risk is problematic. With further conceptual refinement and more extensive empirical support, it may be possible to expand the potential use of these instruments in practice.

At the agency level, resources must be allocated to ensure that risk assessment instruments are being implemented as intended. Consistency and reliability can be improved by training and quality assurance programs (Barber & Delfabbro, 2003; English & Aubin, 1989; English et al., 1992). Although these processes do not address concerns about the validity of risk assessment instruments, they minimize the extent to which validation studies are confounded by the poor reliability of instruments in practice.

References


roundtable on CPS risk assessment (pp. 5–27).


